Capstone Project – Predicting Car Accident Severity

### Alejandro Graizer

1. **Introduction**

Nowadays, and for over 100 years, motor vehicles are one of the most used transport means.

They are used to go to and from work, take children to school, travel, occasional outings, go shopping or simply go for a walk with the family. Given the high use of cars in society, traffic jams and problems are often generated. This is accentuated because, in addition to private cars, the rest of the vehicles that circulate on the same road network must be taken into account: motorcycles, bicycles, buses, electric scooters, trams and even pedestrians.

One of the important factors indicating the development of a city is the state of its road network. That a city has a transit system in good condition, safe and well connected translates into a better quality of life for its inhabitants. Therefore governments invest a large part of their budget in maintaining and improving the road network.

One of the main consequences of poor traffic management is an increase in the number of accidents that occur. This increase becomes a problem, since it not only translates into a health issue for those injured, but also generates an expense for society in terms of both money and human resources.

Because of this, it is crucial to be able to determine the main causes of accidents in order to minimize their severity. There are many sectors interested in creating predictive models that, considering factors such as weather, road conditions, visibility, or type of vehicle, can determine the severity of accidents. Among the interested groups, we can find the government, police departments, insurance companies or road authorities, among others.

These predictive models with their own "Machine Learning" techniques are useful for the government since they could better understand why accidents happen. In this way they can act accordingly and allocate resources for example for better road construction or to pass laws establishing the mandatory use of safety devices such as seat belts.

These models are also attractive to the vehicle manufacturers themselves since having the information about what factors contribute negatively to the severity of the accidents, they can make modifications to counteract it, such as installing airbags.

Road safety authorities may be a possible audience for this type of model. With the insights that can be obtained from these analyses, it is possible to determine where to invest in improving the condition of roads, increasing signage, adding traffic lights, etc.

Another example of who the model could be presented to is car insurance companies. Taking into account the implications of each factor in the severity of accidents, they can improve their business by increasing the premiums paid by some clients, decreasing the deductibles or even not insuring some clients who do not meet minimum requirements.

The aim of this project is to build a model capable of predicting the severity of future accidents in order to reduce future problems originated by this. It will be based on current real data extracted from a database which collects information between the years 2004 and 2020.

1. **Data**

To choose the right database, I reviewed several sources available on the course website. The database titled "Accident risk prediction based on heterogeneous sparse data new dataset and insights" did not seem to me to be suitable for the analysis because the target variable of this source (accident severity) is based on the impact that the accident has on the traffic, and it is not the objective of this project. The second source analyzed was the database "High-resolution road vehicle collision prediction for the city of Montreal", which only showed data from Montreal between 1999 and 2017, which I do not consider representative for my project.

I studied the database provided by the UK police department (https://data.gov.uk/) and discarded it because it only takes into account accidents in which people are injured or killed and does not include accidents of lesser severity in which there is only material damage.

Finally, I decided to use the source provided by the course, the file called "Data-Collisions" (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). Together with the dataset that I chose I have a metadata file explaining the whole variables included. This source is more updated since it covers the years 2004-2020 and contains variables that I consider significant to better explain the severity of the accidents.

This chosen database contains **38 variables** and **194,673 recorded accidents**. Of these variables, I consider that the most relevant would be

- **"Address type"** which allows me to see the type of road in which the accident occurred

- **"Collision type", "Person Count"** to see the type of accident and the people involved

- **Weather", "Road Condition", "Light Condition"** which give circumstantial information of the accident and the road conditions at that time

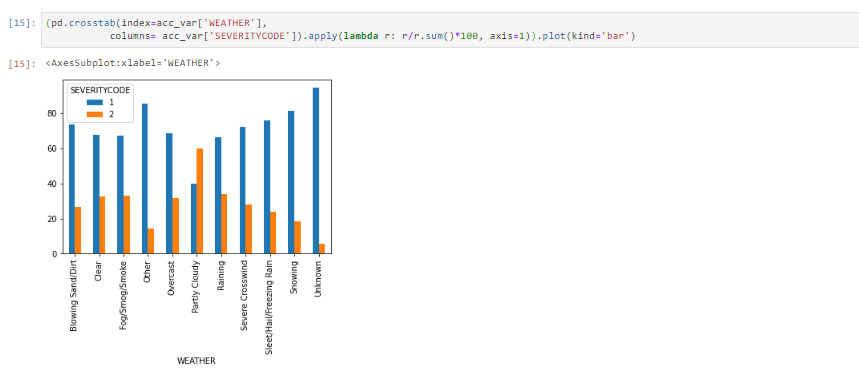
- **"Speeding"** which gives information on whether the speed of the vehicle influenced the severity of the accident

These are the variables that can provide more information to explain the causes of an accident and predict its severity. As an example, if visibility conditions are poor or the vehicle is traveling at high speed, the accident is more likely to be severe.

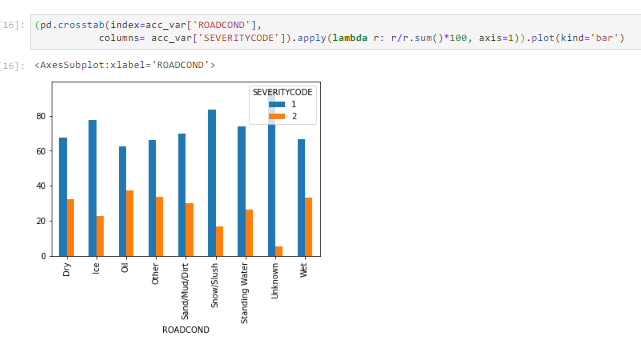
In addition to this, the database has my objective variable **"Severity Code"** that gives information about the severity of the accident and that I will use to check the validity of the model. In this particular case, the variable can take the values 1 (Property Damage) or 2 (Injury). This variable was duplicated in the database, so it has been necessary to remove the repeated variable

Checking the dataset, I saw that there were missing values in different columns. I analyzed which was the best way to treat these fields, whether filling in the values with the averages or with the top or other options. As most of the variables were categorical, the process was not so easy. Seeing that the number of records with empty fields was not significant compared to the number of records I had in my dataset, I decided to eliminate all those rows.

Checking the dataset, I saw that there were missing values in different columns. These allow me to better understand the influence between variables and or their distribution along the different options.



Given that the distribution of accidents of severity 1 and 2 is very similar among all the results of the attribute "WEATHER" I discard this variable as an explanation for our target variable.



Given that the distribution of accidents of severity 1 and 2 is very similar among all the results of the "ROADCOND" attribute, I discard this variable as an explanation for our target variable.



Given that the distribution of accidents of severity 1 and 2 is very similar among all the results of the attribute "SPEEDING" I discard this variable as an explanation for our target variable.

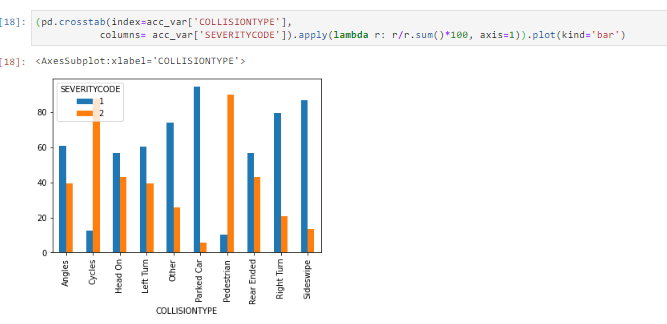
In this way, I form a new dataframe in which I remove the variables that, given the previous checks, are not relevant to our model.

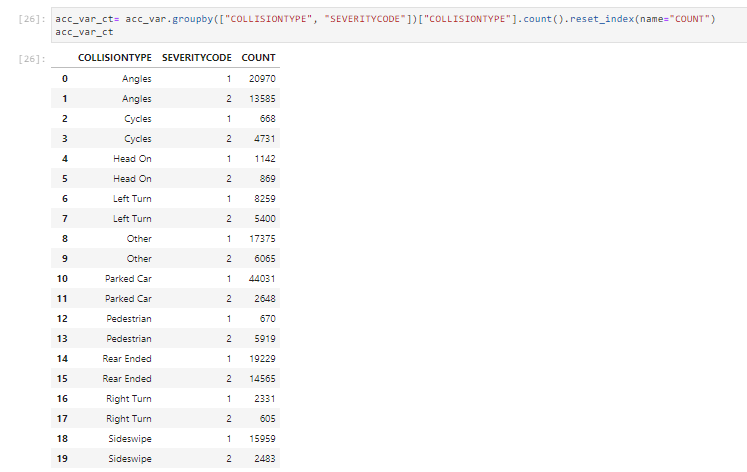
The variables that I will continue analyzing and those that I will use then for the assembly of the predictive model will be:

* CollisionType
* Addrtype
* PersonCount
* SeverityCode (Target variable).

1. **Exploratory Data Analysis**

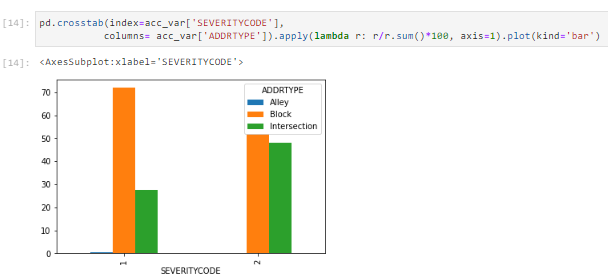
## CollisionType

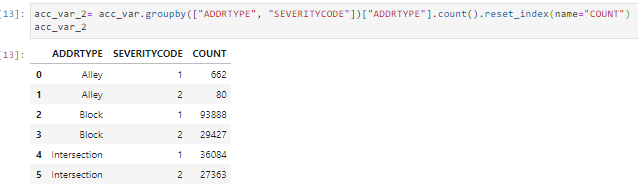




With these numbers and graphics, we can see how the variable "COLLISIONTYPE" influences to determine if the accident is severity 1 or 2. Above all it is evident that there is a certain similarity in this variable along the distribution of the "COLLISIONTYPE" but it is very noticeable how the probability of an accident of severity 2 increases when the "COLLISIONTYPE" includes pedestrians or cyclists.

## ADDRTYPE

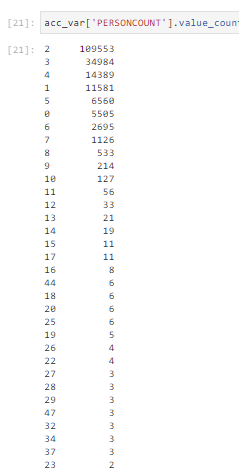




With these numbers and graphics we can see how the variable "ADDRTYPE" influences to determine if the accident is of severity 1 or 2.

Above all it is evident that there is a certain similarity in this variable along the distribution of the "ADDRTYPE" but it is very noticeable how the probability of an accident of severity 2 increases when the "ADDRTYPE" is in an intersection.

## PERSONCOUNT

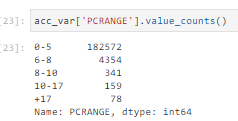
In the variable "PERSONCOUNT" I decided to group the distribution in certain bins. I did this because there were many results in this variable and they were very dispersed. In addition, what happened was that there were certain values in which there were many records and others in which almost none.

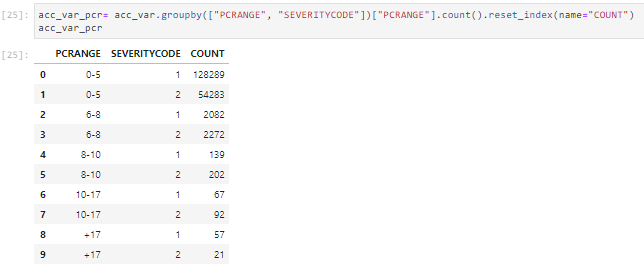
Analyzing the variable by looking at the number of different results became very difficult. Given the large difference in number of records between a result and another, make graphs for visual analysis was impossible because it was impossible to put everything on a single scale.

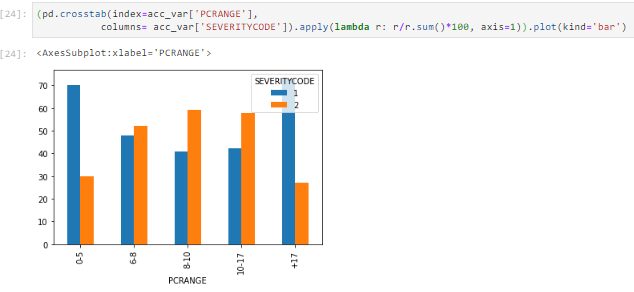
In this way I grouped all the records that had results within the following intervals.

* [0;5]
* [6;8]
* [8;10]
* [10;17]
* +17

I called this new variable with the bins grouping the results "PC RANGE". It is this variable with which I made the tests and graphs to analyze the relevance to predict my target variable and the importance within my model.







With these numbers and graphics, we can see how the variable "PC RANGE" influences to determine if the accident is of severity 1 or 2.

Above all, it is evident that there is a certain similarity in this variable throughout the distribution of the "PC RANGE" but it is very noticeable how the probability of an accident of severity 2 increases when the "PC RANGE" is between 6 and 17 people involved in the accident.

## MODEL LOGISTIC REGRESSION

To make a model that, from certain data of car accidents predicts the objective variable that is the severity of such an accident, I decided that the best thing is to formulate a logistic regression.

Logistic regression is a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) that in its basic form uses a [logistic function](https://en.wikipedia.org/wiki/Logistic_function) to model a [binary](https://en.wikipedia.org/wiki/Binary_variable) [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable). In [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), logistic regression is [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the parameters of a logistic model (a form of [binary regression](https://en.wikipedia.org/wiki/Binary_regression)). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an [indicator variable](https://en.wikipedia.org/wiki/Indicator_variable), where the two values are labeled "0" and "1".

I decided to perform a logistic regression because my case is very similar to the typical characteristics of these. In my case, the target variable is also a binary variable. It only takes as results "1" or "2" and this is an identification of the associated accident. "1" if there was only material damage and "2" if there were injured people.

Other machine learning tools were not suited to the dataset of my case and did not exactly fit the needs of this job. For example, multiple regressions, linear or non-linear, are for predicting a continuous numeric variable.

To do the logistic regression, I had to treat all the categorical variables to transform them into numerical variables. This process consisted in turning each categorical variable into a dummy variable. This means that the column of the variable is replaced by a dataframe in which each column has all the possible options of the variable. Each record is marked with a "1" under the corresponding column and with a "0" under the results that do not correspond. In this way many columns are added to the dataframe on which to run the regression but all fields are numeric (those that were not, now become "0" and "1").

I decided to make my model by separating all my dataset in one part to make the train of the model and another part to make the test of the model. I preferred to take this measure because to make the test with a part of the dataset that was taken to make the train of the model can lead to compete in overfitting errors.

So I took 80% of the data to develop the model (train set) and the remaining 20% of the data I used to test the model and compare the value predicted by the model against the actual value in the database.

To take the samples from the dataset I used the train\_test\_split function that makes these separations from the data in a random way. This is important to do because otherwise you could be leaving out some important pattern within the data.

## Results

I have made 3 times the model with some variants to be able to determine later which one was the best.

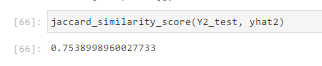
The first one I did was without normalizing the variables and with the "Liblinear" solver.

The second model I made was normalizing the variables and maintaining the same solver.

The third model I made was without normalizing the variables, with the solver "saga" and changing the sample size in the Train\_test\_split function. In this last model I separated 85% for the train process and 15% for the test.

The results have been good and do not vary much from one to another.

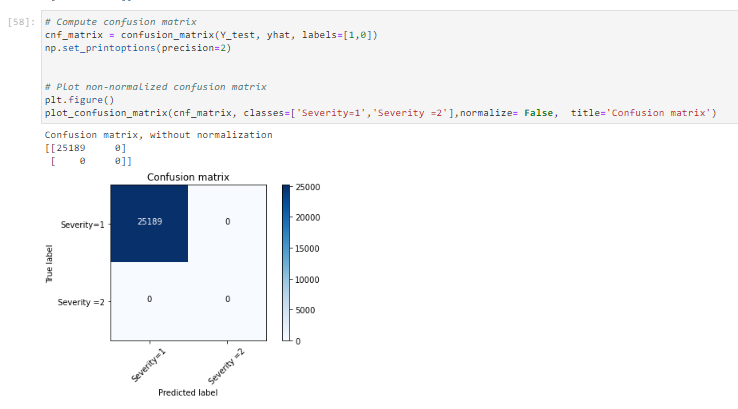
One of the metrics I have used to evaluate the results of the models is the jaccard similarity coefficient score, defined as the size of the intersection divided by the size of the union of two label sets, is used to compare set of predicted labels for a sample to the corresponding set of labels in y\_true . That way I can compare how similar the predicted variable is to the actual variable.

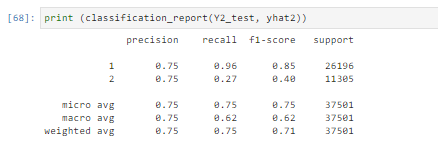


This means that the predicted values match the actual values in the database by 78.39%.

I have also used the confusion matrix to evaluate the models. This is a tool that allows the visualization of the performance of an algorithm that is used in supervised learning. Each column of the matrix represents the number of predictions for each class, while each row represents the instances in the actual class. One of the benefits of confounding matrices is that they make it easier to see if the system is confusing two classes.

Given that in this case, the vast majority of the records are of severity type "1", when you make a subset and perform this confusion matrix, you can see how all the predictions have hit the correct value since they were all of severity 1.





* The recall means "how many of this class you find over the whole number of element of this class"
* The precision will be "how many are correctly classified among that class"
* The f1-score is the harmonic mean between precision & recall
* The support is the number of occurence of the given class in your dataset

Finally, I used the Logloss as a metric to evaluate my model. This metric quantifies the accuracy of a classifier by penalising false classifications. The closer this value is to 0, the better the predictive value of the model. Otherwise, the LogLoss is close to 1.

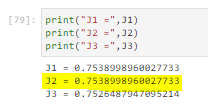


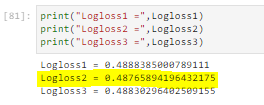
Finally, I make a comparison with the Jaccard similarity Coeficcient Score metric between the 3 different models I formulated.

Similarly, I compared the same models with the LogLoss metric.

With both methods of comparison, the best model is the second one.

With this we can demonstrate that, for this case, the model in which we use the "Liblinear" solver, we normalize the variables is the best. Also, in this sense, we see that a better predictive quality of the target variable is achieved by performing a Split of the dataset with 80% to train the model and 20% to test than 85% and 15% respectively.





## DECISIONS

After these results, the hearing of this report should make us aware of why and in what situations these accidents are occurring.

The government, the Police, the road authorities have to start taking resignations from when these accidents occur and under what circumstances these accidents are most likely to be serious or not. Clearly, after this report the authorities have to take into account that, for example, it is in the road intersections that a serious accident is most likely to occur. It is there, then, that they should take extreme precautionary measures and ensure that everything follows the rules so that the number of events is reduced.

It is also very likely that there will be serious accidents, when there are many people involved (more than 6). It is in this way that those companies that are dedicated to the manufacture of large vehicles, in which more people enter, should review their processes, improve their products to reduce the number of accidents or at least ensure that when they happen, there are no injuries between people (not accidents of severity 2).

On the other hand, road authorities, after this report, may know that in all likelihood, accidents involving pedestrians or cyclists are very likely to be serious. They need to be able to act on this new knowledge so that roads on which the connection of cars with pedestrians or bicyclists is endangered are better cared for.

## CONCLUSIONS

In this study I have analyzed the relationship between the number of people in an accident, the type of road and the type of accident with the severity of such event. This report can serve as a basis for taking action on a city's road system. It can be seen how the number of accidents that occur in a city greatly affects the lives of its inhabitants. It is a high priority to make all possible changes so that they do not occur or are at least not as severe.

In the same vein, I think it is important that governments begin to realize how useful these analyses can be and to encourage them. The best way to do this could be to commission a data collection team or department within each city. I have done this report in general for all cities but it can be very valuable to have more data, with more characteristics about accidents. For example, to know the location, if there was a responsible, age of the people involved, time of the accident and other specific characteristics of each City so that when making the studies more precise measures can be taken of each situation.